Similarity between world capitals with foursquare data

Introduction

Globalization is a natural event that has been accentuated in recent years. We see more and more cultures transposing and cities having more similarities. Nowadays we find Japanese, Mexican and Arab restaurants in a large number of world capitals.

What other establishments are hallmarks of the globalization phenomenon? Which cities are most similar to each other?

Methodology

Data

The list of the world capitals locations is obtained at https://lab.lmnixon.org/4th/worldcapitals.html, it is composed of 200 entries and the locations of the cities come in longitude and latitudes.

The locations in each city come from Foursquare. The acquisition is made through the Foursquare API, which finds the venues within the radius of the informed location. Along with the number of locations found, the API provides the categories of locations, for example: hotel, cafe, historic site, Asian restaurant, etc. In this work, the chosen radius is 100,000. The goal with this large number is to get a city-sized radius and thus be able to get the most samples from locations to get a good picture of facilities across the city.

Model

The goal here is to clustered the cities in similarity of venues. For that it is used a K-means algorithm with the hyper parameters: n_cluster = 5, random_state = 0, n_init = 10, max_iter = 300, tol = 0.0001, precompute_distances = 'auto', verbose = 0, random_state = None, copy_x = True, n_jobs = None.

Results
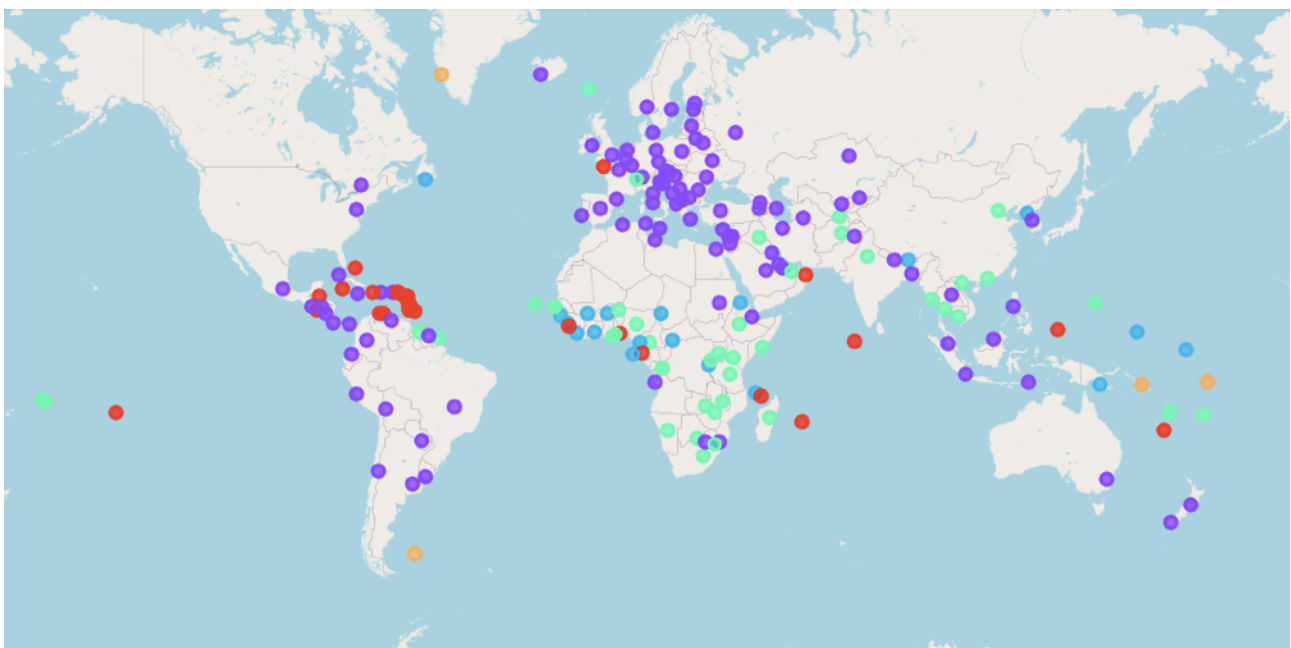
- The purple cluster consists of 102 capitals and the 5 most common venues, in order, are parks, hotels, cafés, restaurant and coffee shops.

- The blue cluster consists of 19 capitals and the 5 most common venues, in order, are hotels, resorts, airports, farms and falafel restaurants.

- The green cluster consists of 44 capitals and the 5 most common venues, in order, are hotels, cafés, shopping malls, restaurants and coffee shops.

- The orange cluster consists of 4 capitals and the 5 most common venues, in order, are café, airports, hotels, fast food restaurants and entertainment service.

- The red cluster consists of 28 capitals and the 5 most common venues, in order, are beaches, hotels, bakery, restaurant and coffee shops.

Discussion

We can see a strong presence of establishments related to leisure and tourism. This made it difficult to separate the groups. A suggestion for further analysis to solve this problem is to eliminate the most common establishments, stay with those that really characterize a place.

Looking further at this challenge, two hypotheses arose as to why this database presents this similarity between the entries. If it is a bias brought by Foursquare, it may be because Foursquare has become strong within the leisure and tourism niche. Now, if it is due to the moment of history of these cities that the data were collected, we can imagine that in the stage of globalization in which we find ourselves in the 21st century, these are expected to be the most common establishments. To test these hypotheses, further analysis is needed.